N91-24045

ELECTRONIC NEUROPROCESSORS

Anil Thakoor

Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, California 91109

ABSTRACT

Jet Propulsion Laboratory's Center - for Space Microelectronics Technology (CSMT) is actively pursuing research in the neural network theory, algorithms, and electronic as well as optoelectronic neural net hardware implementations, to explore their unique strengths and application potential for a variety of NASA, DoD, as well as commercial application problems, where conventional computing techniques are extremely time-consuming, cumbersome, or simply non-existent. This paper presents an overview of the JPL's electronic neural network hardware development activities and some of the striking applications of the JPL:s electronic neuroprocessors.

INTRODUCTION

You enter a crowded room and instantly recognize a familiar face at the far end. You wave, smile, and effortlessly weave your way through the people and furniture to reach your friend. You neither go in a straight line, nor follow a well calculated path with the least "cost" of traversing, but you proceed swiftly and efficiently anyway. In fact, your mind is already racing through the memories of a tennis game you played with that friend over ten years ago!

Digital computers today allow us to plan extremely complex, multiyear, deep space missions with amazing accuracy and provide us with enormous computing power; but something as simple to us as recognizing a face or "a pattern" in a cluttered background is not at all easy for even today's state-of-the-art supercomputers. Elaborate "expert" systems based on collective knowledge of many experts, arranged systematically to form rule bases, provide excellent software tools as "artificial intelligence". However, there is no convenient way yet to really "capture" in a computer the unique skills for example of a veteran fighter pilot with years of experience, such that they can be easily analyzed and transferred to others. The best advice from a maestro would be: "just watch how I do it and learn", something we may consider doing ... but a hopeless proposition for today's computer.

The emerging field of artificial neural networks[1], inspired by the functioning of a human brain, attempts to capture some of its unique abilities in learning, self-organizing, and intelligent information processing at extremely high speeds even with fuzzy inputs and ill-defined situations, to complement the powerful, high accuracy number-crunching digital machines. The secret of biological neural networks lies in their complex, massively parallel architectures. A human brain consists of over 10 billion neuron cells communicating among themselves through networks of over 100 trillion synaptic interconnections! Even though many of the intricacies of the brain functions are far from well-understood, it is recognized that massive parallelism, distributive storage, and a synchronous, analog, concurrent processing of information are some of its key attributes. A variety of architectural models and neural net algorithms have emerged during the past several years, through extensive software simulations, with a primary objective of developing better understanding of the unique capabilities of the biological machines and a secondary goal of capturing some of those attributes in computer systems. However, the potential of high speed from the massively parallel processing by "artificial" neural nets, mimicking the architectures of their biological counterparts, will be realized only when the architectures are actually implemented in parallel hardware[2]. Can it be done? How can one build large arrays of artificial neurons and synapses and orchestrate their simultaneous operations without a system clock, as the nature does it? What are the best suited technologies and device structures? And finally, which real-world problems could substantially benefit today from the neuro-processing approach?

JPL'S APPROACH

Over the past several years, JPL's Center for Space Microelectronics Technology has sought answers to several of these questions through multifaceted research programs on theory, algorithms, hardware implementations, and applications of artificial neural networks. In particular, JPL has pioneered the development of fully parallel analog hardware implementations of neural network architectures in electronics[3-7], to understand the dynamics of such massively parallel nonlinear systems, as well as to apply the new powerful computing paradigms to problems not solved easily by other techniques.

The basic components of electronic neural network hardware are conceptually and functionally extremely simple; the neurons can be implemented as thresholding nonlinear amplifiers, and the synapses as variable resistive connections between them[5]. An artificial neural network therefore consists of many simple processing elements or tailored amplifiers, representing neurons, which interact among themselves through networks of weighted connections functioning as synapses. The computation performed by the network is primarily determined by the network topography and the synaptic weights. The state of the system is identified by the pattern of activity of the neurons. Given an initial activity pattern, each neuron receives input signals from other neurons and adjusts its output accordingly over time. The system rapidly evolves into a steady activity pattern which is then interpreted as a memory recall or as a solution to a problem.

Several promising neural network architectures developed over the years utilize two broad classes of connection schemes: a fully-connected feedback architecture and a multi-layered feedforward structure, illustrated in Fig. l(a) and (b), respectively. In a feedforward network, neurons from each layer broadcast their outputs only to the neurons of the subsequent layer, modulated by the synaptic weights. The information processing thus progresses in the forward direction. In a fully connected feedback network on the other hand, all the neurons can interact dynamically with one another in parallel through the feedback synapses. Thus, the dynamics of feedback networks plays an especially important role in dictating their emergent computational properties.

The most important operational characteristic of such architectures, however, remains their massive parallelism. The highly distributed and effectively redundant information collection, storage, and manipulation in the multitude of synaptic weights give rise to inherent fault tolerance resulting in graceful degradation in their performance. On the other hand, the concurrent analog processing by a large number of neurons when implemented in hardware promises computing speeds orders of magnitude higher than serial processors. Above all, the highly parallel neural network algorithms provide unique abilities to solve computation-intensive global optimization problems and to "learn" fuzzy transformations from examples in ill-defined situations.

To be able to fabricate a variety of neural network architectures in a fully parallel fashion by using only a few selected, generic, "building blocks", JPL has developed, designed, and fabricated two separate families of cascadable custom-VLSI chips in analog CMOS: two dimensional arrays of fully programmable synapses, and one dimensional arrays of non-linear neurons.

HARDWARE IMPLEMENTATIONS

JPL's reconfigurable building blocks include a spectrum of cascadable, programmable, synaptic and multi-neuron chips (Fig. 2) with tailored functional characteristics. A typical synaptic chip consists of a fully connected 32 X 32 array of synaptic devices fabricated using the standard 2 micron bulk CMOS process. The synaptic connection embodiments vary from a simple binary (ON/OFF) scheme to fully parallel analog designs exceeding 10-bit dynamic resolution. A variety of methods have been used to obtain variable synaptic weights on the VLSI chips. For example, simple long-channel CMOS transistors provide programmable synaptic weights with binary values, "on-chip" static memories and multiplying digital to analog convertors (MDAC) furnish synapses with weight resolution of up to 7 bits, and four-quadrant multipliers that scale connection strengths according to voltages residing on invisibly charge-refreshed capacitors result in weight resolution exceeding 10 bits, a major achievement, particularly important for our currently ongoing investigations on supervised and unsupervised learning in neural network hardware. Our cascadable neuron

array chips (Fig. 2) provide a unique "variable gain" feature that is valuable for embodiments of networks of varying sizes, as well as to generate controlled "simulated" annealing required during convergence of feedback networks. Such fully programmable building blocks provide a very convenient library of hardware to construct suitable network architectures dictated by a problem.

Compact hardware implementations of such massively parallel architectures of course differ significantly from conventional digital designs. For example, due to the overall power dissipation concerns in the parallel circuitry, the synaptic connections modulating communications among the neurons need to be extremely "weak" or highly resistive. Such unusual requirements of the massively parallel and nonlinear processing in neural net hardware present a totally new set of interesting issues regarding precision and tolerance, influence of static and dynamic noise sources, and the useful dynamic resolution in the analog information being processed. The synaptic connection elements on a CMOS chip for example utilize high-precision, long channel field effect transistors, providing weak, current-limiting connections in their fully "ON" state and several orders of magnitude higher resistance in the "OFF" state. Furthermore, resistivity-tailored thin film elements of cermets integrated with the memory-switching devices promise high density ($\sim 10^8$ synapses/cm^2) for the synaptic arrays.

NOVEL DEVICE STRUCTURES

Although conventional VLSI technology offers a convenient approach to implementing "neuro-functions" in hardware, its circuit concepts tailored for sequential processing result in unnecessary hardware complexity to accomplish simple functions. For example, a programmable analog synaptic connection, essentially a resistive component, requires a large number of transistors occupying expensive silicon area and thus limits the size of the implemented network. Ideally, a two-terminal programmable solid state memory device would simplify large network implementations significantly. For the next generation neural network hardware therefore, JPL is investigating several novel thin film device structures based on materials with tailored electronic properties and analog thin film-VLSI hybrid device concepts. JPL's thin film device efforts have already demonstrated programmable nonvolatile synapses based on memory switching in hydrogenated amorphous silicon and manganese oxide with a potential to realize extremely high synaptic density, approaching 10^9 synapses/cm^2, suited for a variety of massive data management applications. Furthermore, optically addressable, analog memory devices based on ferroelectric thin films are also under development for large, 2-dimensional (focal plane) synaptic arrays for optoelectronic implementations of neural networks.

NEUROPROCESSOR APPLICATIONS

Unique strengths of neural networks complement the power of conventional digital computers very well. Hardware implementations of tailored neural net architectures therefore become extremely high speed, special-purpose, function-specific "co-processors" interfaced to digital computers. JPL is heavily involved in developing such application-specific neuro-processors for complex problems, where digital techniques are either limited in scope and speed or simply not applicable due to the computation-intensive and fuzzy nature of the problems. Figure 3 shows an example of a VLSI neuroprocessor interfaced to a personal computer. The neural network in this case is processing several analog constraints (e. g. conditions of soil, surface roughness, slope, vegetation, and rain) simultaneously, to determine the "cross-country mobility" for a military vehicle over a terrain under consideration, with a manyfold speed enhancement compared to the digital machine alone.

Another striking example of the enabling nature of the neuroprocessing approach is evident from our dedicated neuroprocessor for resource allocation, currently under development. Based on our innovative "analog prompt scheme" and "limited connectivity" architecture already demonstrated in hardware, the resource allocation processor promises real-time solutions to computation-intensive problems of global optimization and dynamic assignment such as pairing of resources to consumers (or assignment of weapons to targets) to minimize the "global cost" involved in the situation. For example, a 64 resource to 64 consumer assignment problem, for a one-to-one association, involves a cost matrix of 64 X 64, and a total of 64! or 10^{89} possible solutions. For a problem of this size, the neuroprocessor promises an optimal or near-optimal

solution within a fraction of a millisecond. This is over three orders of magnitude faster compared to conventional heuristic search techniques, even when running on multiprocessor machines such as hypercube. Moreover, the reconfigurable neuroprocessor offers solutions to dynamic assignment problems even with arbitrary many-to-many association constraints, extremely difficult for digital computing methods.

Other neuroprocessors developed by JPL are under evaluation at present for applications in cartographic analysis, terrain feature recognition from landsat imagery, and best path determination in a constrained space. Furthermore, JPL has developed one of the first ever reconfigurable, multilayer neuroprocessor systems with a capability of learning in analog hardware. This system is currently applied to problems of computation-intensive inverse kinematic transformations in robotics and ill-defined feature recognition from multispectral images.

Clearly, the strength of neuroprocessors is complementary to the digital computers. To combine the best of both worlds, The future supercomputers with multiple digital processing nodes (e. g. hypercube architecture), may have several special purpose neuroprocessors "servicing" individual computing nodes. Such a system may also have additional analog neuroprocessors for example to carry out specialized tasks such as load balancing and problem decomposition, where neural network-derived methods show great promise.

CONCLUSIONS

JPL's fully parallel, programmable, neural network hardware has not only provided high speed research tools to investigate unique emergent computational properties of neural networks, but has also furnished the "building blocks" for development of special-purpose, application-specific neuroprocessors. Powerful artificial neural networks have been implemented in fully parallel hardware, in spite of the inherent peculiarities and unavoidable noise constraints, characteristic of analog hardware. Analog, parallel neuroprocessors provide orders of magnitude speed enhancement and/or totally new capabilities compared to conventional digital techniques.

ACKNOWLEDGEMENTS

The work described in this paper was performed by the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored in parts by the Defense Advanced research Projects agency, the Joint Tactical Fusion Program Office, and the Strategic Defense Initiative Organization/Innovative Science and Technology, through an agreement with the National Aeronautics and Space Administration.

REFERENCES

1. McClelland, J. L.; and Rumelhart, D. E. : Parallel Distributed Processing. Part I and II. The MITPress, (Cambridge, MA), 1987.

2. Mead, C. : Analog VLSI and Neural Systems. Addison Wesley, USA, 1989.

3. Thakoor, A. P.; Moopenn, A.; Lambe, J.; and Khanna, S. K. : Electronic Hardware Implementations of Neural Networks. Applied Optics, vol. 26, no. 23, 5085, 1987.

4. Eberhardt, S.; Duong, T.; and Thakoor, A. P. : Design of Parallel Hardware Neural Network Systems from Custom Analog VLSI Building Block Chips. Proc. IEEE/INNS Int'l Joint Conf. on Neural Networks, vol. 2, 183, 1989.

5. Moopenn, A.; Lambe, J.; and Thakoor, A. P. : Electronic implementation of associative memories based on neural network model. IEEE Trans. Sys., Man, and Cyber., SMC-17, 325, 1987.

6. Daud, T.; Moopenn, A.; Lamb, J. L.; Ramesham, R.; and Thakoor, A. P. : Neural Network Based Feed-Forward High Density Associative Memory. Proc. IEDM, 107, 1987.

7. Eberhardt, S.; Duong, T.; and Thakoor, A. P. : A VLSI Analog Synapse Building Block Chip for Hardware Neural Network Implementations. Proc. Third Annual Parallel Processing Symposium, Vol.1, 257, 1989.

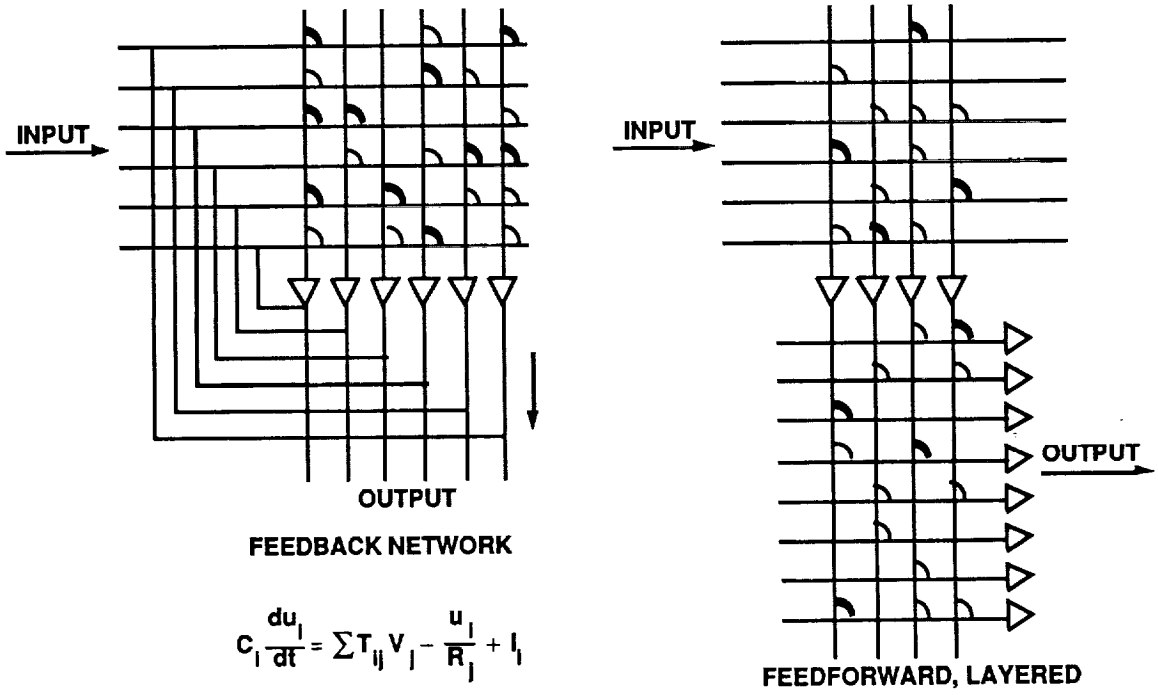

Figure 1. Schematic diagram of (a) a fully connected feedback neural network and (b) a multilayer feed-forward neural network architecture.

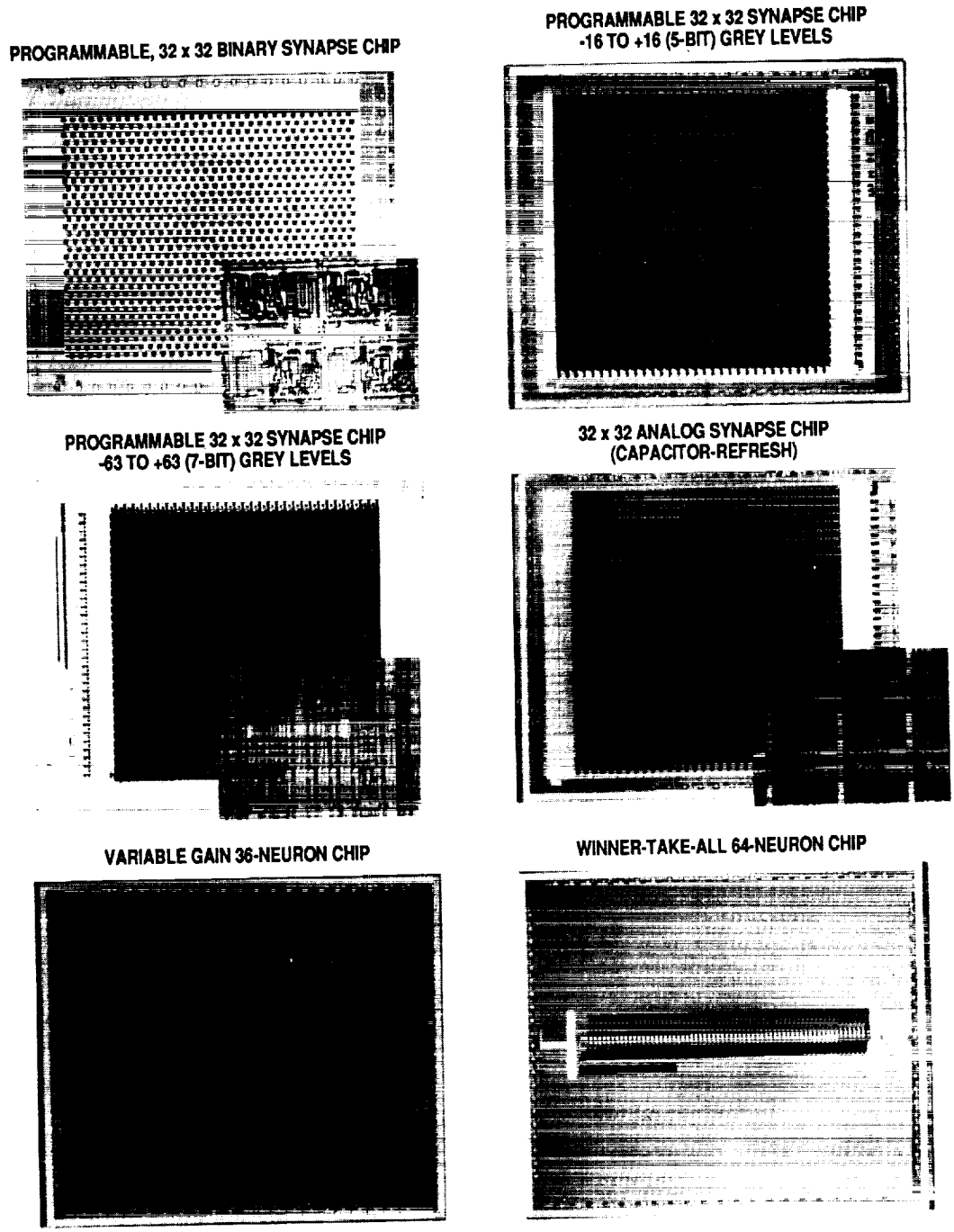

Figure 2. Cascadable custom-VLSI neural network building block chips fabricated using analog CMOS technology.

Figure 3. A neuroprocessor interfaced to a digital computer, configured for terrain trafficability determination.